BMNB 852: Applied Bioinformatics

Week 6, Lecture 11

István Albert

Bioinformatics Consulting Center
Penn State, 2015
String similarity - edit distances

The number of changes to transform from one string to another (based on some rules).

LOVE $\rightarrow$ HATE

Edit distance of 3 (replace L, replace O, replace V)

If we were to impose a vocabulary (all intermediate words need to be nouns) the edit distance may change:

LOVE $\rightarrow$ ? $\rightarrow$ ? $\rightarrow$ ? $\rightarrow$ HATE

(also know as the game called: Word Ladder invented by Lewis Carroll)

String similarity can be computed in difference contexts!
Sequence Alignments

• A way to arrange sequences to identify regions of similarity

ATGCAAAACAAG

ATGCTTATTAG

• We look for similarity because it may be a consequence of functional, structural or evolutionary relationships
How to pick the right alignment?

We need a scoring scheme:
- positive values to reward matches
- negative values to penalize mismatches
Scoring Alignments

Values are associated to:

- an exact match → usually a positive score (5)
- a mismatch → usually a negative score, may depend on the kind of mismatch (-4)
- gap opening → usually the most penalized action (-10)
- gap extensions → making the gap longer (-0.5)

These are called **scoring matrix**

**Important to remember**

1. **Any two sequences can be aligned**, the alignment score represents the sum of the each match/mismatch/gap/gap extension

2. There is no such thing as universally best alignment only the **optimal alignment for a given scoring matrix**

3. Most aligners will **only report alignments that make some sense** and usually the longest alignment in a region (this is not that simple as it sounds)

4. Alignment is a measure of **similarity** but not homology (shared ancestry)
Nucleotide scoring matrix

```
# This matrix was created by Todd Lowe 12/10/92
#
# Uses ambiguous nucleotide codes, probabilities nearest integer
#
# Lowest score = -4, Highest score = 5
#
<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>T</th>
<th>G</th>
<th>C</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>5</td>
<td>-4</td>
<td>-4</td>
<td>-4</td>
</tr>
<tr>
<td>T</td>
<td>-4</td>
<td>5</td>
<td>-4</td>
<td>-4</td>
</tr>
<tr>
<td>G</td>
<td>-4</td>
<td>-4</td>
<td>5</td>
<td>-4</td>
</tr>
<tr>
<td>C</td>
<td>-4</td>
<td>-4</td>
<td>-4</td>
<td>5</td>
</tr>
<tr>
<td>S</td>
<td>-4</td>
<td>-4</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>W</td>
<td>1</td>
<td>1</td>
<td>-4</td>
<td>-4</td>
</tr>
<tr>
<td>R</td>
<td>1</td>
<td>-4</td>
<td>1</td>
<td>-4</td>
</tr>
<tr>
<td>Y</td>
<td>-4</td>
<td>1</td>
<td>-4</td>
<td>1</td>
</tr>
<tr>
<td>K</td>
<td>-4</td>
<td>1</td>
<td>1</td>
<td>-4</td>
</tr>
<tr>
<td>M</td>
<td>1</td>
<td>-4</td>
<td>-4</td>
<td>1</td>
</tr>
<tr>
<td>B</td>
<td>-4</td>
<td>-1</td>
<td>-1</td>
<td>-1</td>
</tr>
<tr>
<td>V</td>
<td>-1</td>
<td>-4</td>
<td>-1</td>
<td>-1</td>
</tr>
<tr>
<td>H</td>
<td>-1</td>
<td>-1</td>
<td>-4</td>
<td>-1</td>
</tr>
<tr>
<td>D</td>
<td>-1</td>
<td>-1</td>
<td>-1</td>
<td>-4</td>
</tr>
<tr>
<td>N</td>
<td>-2</td>
<td>-2</td>
<td>-2</td>
<td>-2</td>
</tr>
</tbody>
</table>
```

BLOSUM: BLOck Substitution Matrix (it has versions based on clustering similarity)

```
# Matrix made by matblas from blos
# * column uses minimum score
# BLOSUM Clustered Scoring Matrix
# Blocks Database = /data/blocks_5
# Cluster Percentage: >= 62
# Entropy = 0.6979, Expected =

A 4 -1 -2 -2 0 -1 -1 0 -2 -1 -1
R -1 5 0 -2 -3 1 0 -2 0 -3 -2
N -2 0 6 1 -3 0 0 0 1 -3 -3
D -2 -2 1 6 -3 0 2 -1 -1 -3 -4
C 0 -3 -3 -3 9 -3 -4 -3 -3 -1 -1
```
### Scoring Alignments

<table>
<thead>
<tr>
<th></th>
<th>ACGCA</th>
<th>ATGCA</th>
<th>ATGTA</th>
<th>ATCGAA</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>3</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>4</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>5</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>6</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>7</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

**Scores:**

<p>| | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>7</td>
<td>16</td>
<td>16</td>
<td>15</td>
<td></td>
</tr>
</tbody>
</table>
More on alignment scoring

• It is possible to pick meaningless scoring parameters that will lead to meaningless alignments.

• In the majority use cases we leave them on defaults or use a known matrix (BLOSUM, PAM)

• Overall the scoring matrix should be negative (penalizes errors) otherwise may produce many spurious alignments

• End gaps may/are/should be treated differently than internal gaps (there is an evolutionary rationale for this)
Global alignments: every residue will be aligned

Local alignments: similar regions will be aligned

Global-Local (semi-global) alignments:

- No penalty for gaps at the end
- Aligns one (usually the shorter) sequence fully against the other (longer) one
Global Alignment

Input Sequences

Sequence A

Sequence B

Global Alignment

Sequence A

Sequence B

GAPS

Sequence B

GAPS
Local Alignment

The aligner may produce just the best or a top hits ordered by score.
Alignment Algorithms

- Can be **optimal** (mathematically precise) or **near-optimal** algorithms.

- Optimal alignments are usually computationally VERY demanding. We almost always use near-optimal aligners.

- Optimal global alignment: Needleman-Wunsch, 1970
- Optimal local alignment: Smith-Waterman, 1981

Active area of computer science research new algorithms

- Suffix trees, Suffix array, Burrow-Wheeler Transform, FM index → new bioinformatics applications
Let’s align protein sequences

First sequence:

>1
THISLINE

Second sequence:

>2
ISALIGNED

Cute example taken from the book Understanding Bioinformatics (see course webpage for link)
**Pairwise Sequence Alignment**

Pairwise Sequence Alignment is used to identify regions of similarity that may indicate functional, structural and/or evolutionary relationships between two biological sequences (protein or nucleic acid).

By contrast, Multiple Sequence Alignment (MSA) is the alignment of three or more biological sequences of similar length. From the output of MSA applications, homology can be inferred and the evolutionary relationship between the sequences studied.

**Global Alignment**

Global alignment tools create an end-to-end alignment of the sequences to be aligned. There are separate forms for protein or nucleotide sequences.

**Needle**

EMBOSS Needle creates an optimal global alignment of two sequences using the Needleman-Wunsch algorithm.

**Stretcher**

EMBOSS Stretcher uses a modification of the Needleman-Wunsch algorithm that allows larger sequences to be globally aligned.

**Local Alignment**

Local alignment tools find one, or more, alignments describing the most similar region(s) within the sequences to be aligned. There are separate forms for protein or nucleotide sequences.

**Water**

EMBOSS Water uses the Smith-Waterman algorithm (modified for speed enhancements) to calculate the local alignment of two sequences.

**Genomic Alignment**

Genomic alignment tools concentrate on DNA (or to DNA) alignments while accounting for characteristics present in genomic data.

**Wise2DBA**

Wise2DBA (DNA Block Aligner) aligns two sequences under the assumption that the sequences share a number of colinear blocks of conservation separated by potentially large and varied lengths of DNA in the two sequences.

**GeneWise**

GeneWise compares a protein sequence to a genomic DNA sequence, allowing for introns and frameshifting errors.

**PromoterWise**

PromoterWise compares two DNA sequences allowing for inversions and translocations, ideal for promoters.
Global Alignment

Use the **needle** program.

Usually Gaps: - or spaces
Exact matches: |
Other format specific qualifiers: . : 

Check the documentation for exact details

```
#=======================================
#
# Aligned_sequences: 2
# 1: 1
# 2: 2
# Matrix: EBLOSUM70
# Gap_penalty: 10.0
# Extend_penalty: 0.5
#
# Length: 11
# Identity:  4/11 (36.4%)
# Similarity: 5/11 (45.5%)
# Gaps:  5/11 (45.5%)
# Score:  8.5
#
#
#=======================================

1   1 THISLI--NE--  8
    ||.:  ||

2   1 --ISALIGNED  9
```

#=======================================
#=======================================
Local Alignments

Use the `water` program.

Usually Gaps: - or spaces
Exact matches: |
Other format specific qualifiers: . :

Check the documentation for exact details

```plaintext
#---------------------------------------
#
# Aligned_sequences: 2
# 1: 1
# 2: 2
# Matrix: EBLOSUM62
# Gap_penalty: 10.0
# Extend_penalty: 0.5
#
# Length: 2
# Identity: 2/2 (100.0%)
# Similarity: 2/2 (100.0%)
# Gaps: 0/2 ( 0.0%)
# Score: 11.0
#
#
#---------------------------------------

1          7   NE   8
      ||
2          7   NE   8
```

#---------------------------------------
#---------------------------------------
The scoring determines the alignment

Why did we get this?

NE    IS-LI-NE
||    || || ||
NE    ISALIGNE

Scores:

11    7
# BLOSUM 90

# Aligned_sequences: 2
# 1: 1
# 2: 2
# Matrix: EBLLOSUM90
# Gap_penalty: 10.0
# Extend_penalty: 0.5
# Length: 6
# Identity: 4/6 (66.7%)
# Similarity: 5/6 (83.3%)
# Gaps: 1/6 (16.7%)
# Score: 14.0

1  4 SLI-NE  8
   \| | \| |
2  3 ALIGNE  8
# BLOSUM30

# Aligned_sequences: 2
# 1: 1
# 2: 2
# Matrix: EBLOSUM30
# Gap_penalty: 10.0
# Extend_penalty: 0.5

# Length: 4
# Identity: 2/4 (50.0%)
# Similarity: 3/4 (75.0%)
# Gaps: 0/4 (0.0%)
# Score: 15.0

1
5 LINE 8
:.|||
2
5 IGNE 8
Homework 11 (part 1)

You may use the EMBOSS web or command line interfaces.

Globally align the THISLINE and ISALIGNED sequences with the BLOSUM30 substitution matrix.

- Verify the correctness of the percent identity values.
- Verify (show math) on how the alignment score is computed.
Find the parameters that could lead to a local alignment that looks like this.